Bioinformatics 2, IT and health

Tejal Joshi, Post doc
Center for Biological Sequence Analysis, DTU,
Building 208. Room 61

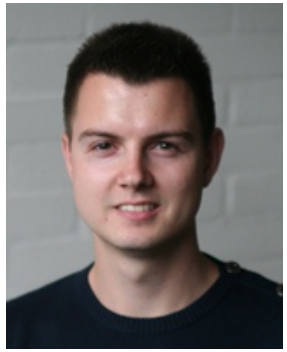
tejal [@] cbs [.] dtu [.] dk

Teachers

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Tejal

Andrea



Christian

Ole

Kristoffer

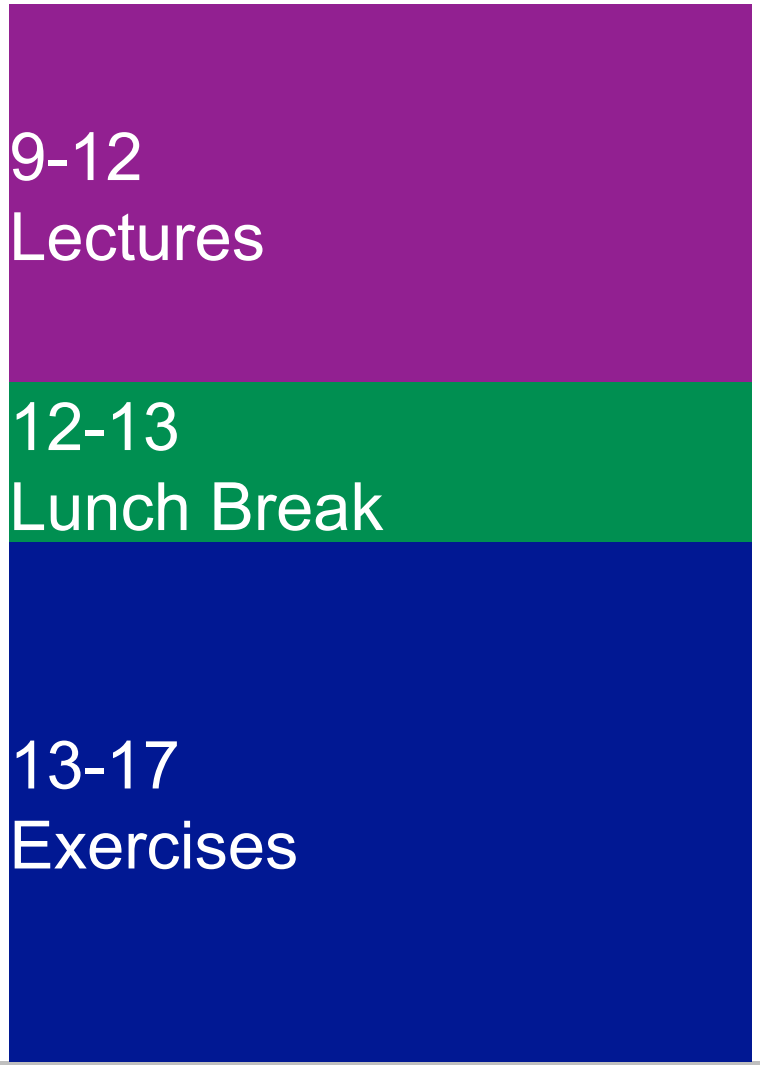
Paolo

Kasper

Technical University of Denmark - DTU
Department of systems biology

DTU

A typical course day



Eduroam

WIRELESS

Setting up your laptop

http://www.cbs.dtu.dk/courses/coursesoftware.php

PRE-COURSE SURVEY RESULTS

Course Program

CENTERFO
RBIOLOGI
CALSEQU
ENCEANA
LYSIS CBS

Tentative schedule:
Day 1. Introduction to Unix system commands, shell scripts, regular expressions.
Day 2. Intro to R programming, statistics and data visualization
Day 3. Python for bioinformatics
Quick refresher on Python, bioPython modules - installations and usage thereof, file parsers (fasta, fastq, gff, blast output, etc file types), plotting simple graphs (using R)
Day 4. Webservices to access remote databases
Day 5. Machine learning
Machine learning concepts intro.,Classification, modeling and predictions. Performance evaluation of classifiers/predictors (sensitivity, specificity, MCC, ROC, AUC, etc).
Day 6. Text mining
Day 7. Machine learning (continued). ANNs, design and evaluation,
Day 8. Summing it all up + discussion on assignments/projects
More information on this course can be found here: http://wiki.bio.dtu.dk/teaching/index.php/27634_Program_November_2013

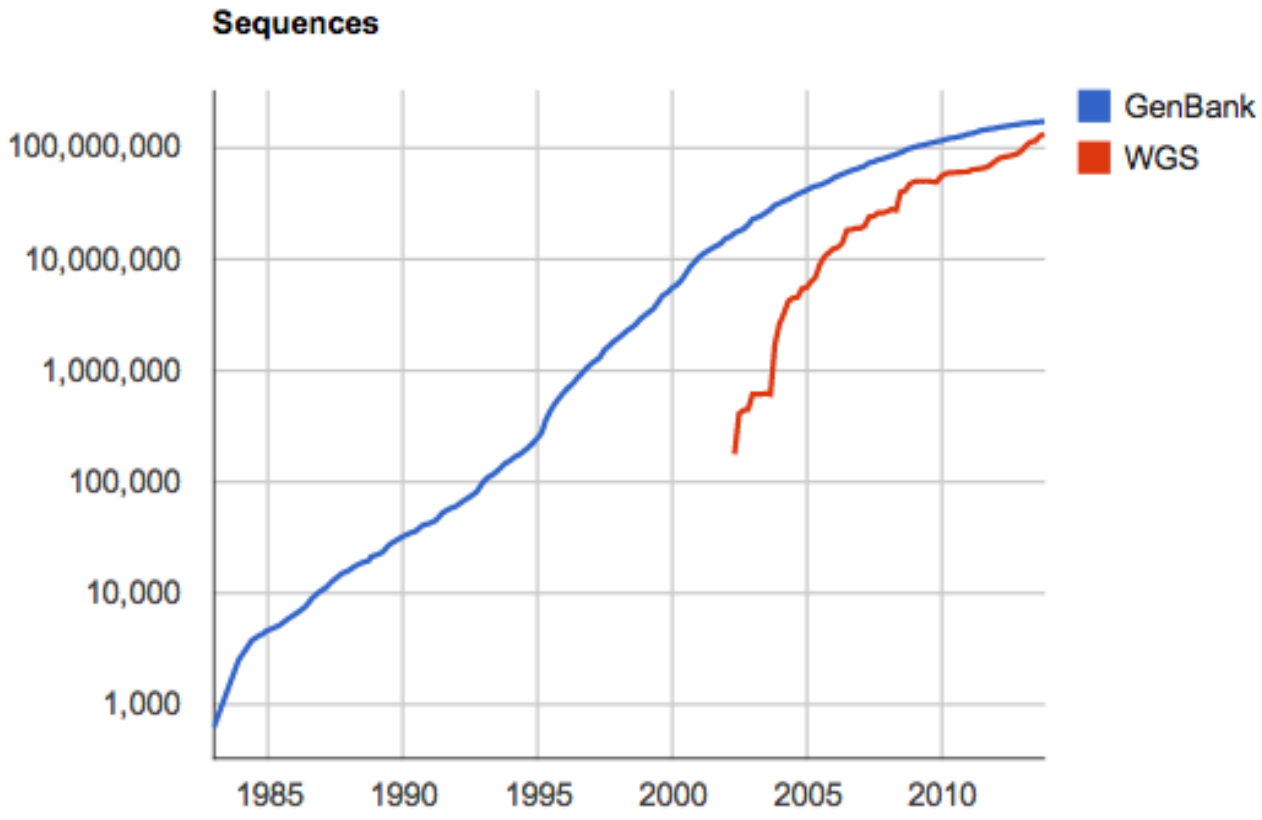
What we aim to teach in the course

- Multiple ways of accessing biological databases
  - Biopython
  - Web services (SOAP, REST)
- Working with data obtained from databases:
  - Format files as per your requirements (Unix/R)
  - How does the data look like ? (Visualization using R)
  - Can we do some basic statistics? (using R)
- Can we learn from what is available and apply the knowledge to learn more using computers?
  - Machine learning (Train the computer model to learn using, for example, artificial neural networks, SVM, etc.)
  - How can the learning be evaluated ? (cross-validation accuracy, sensitivity, specificity, ROC, MCC, and other abbreviations)
- But the information to learn from is often sparse and not readily available
  - Can we derive high-quality information from text ? (Text mining)

Biological Databases

Big data problem

155176494699 bases in Genbank. (October 2013).



Databases

- In general, databases facilitate efficient storage, manipulation and retrieval of information
- Databases are composed of tables of data
- Tables are data structures for logically related sets of data called “records”.
- Each record contains several attributes of information. For example, a student record.
- One can identify or retrieve records based on a unique attribute – called primary key.

Flat file format databases

- Commonly used in biology
- Allows use by multiple applications without requiring complex data types and formats.
- Data stored as plain text (alphabets and/or numbers) with specific delimiter.
- They could be organism specific, data type specific, etc.

Can you give me an example of flat file database in biology ?

FASTA file.

Question: what is the delimiter used ??

Steps in accessing data

- Formulate a query.
  - Retrieve mRNA sequence of gene “PTEN” in human ?
  - Retrieve 3’ UTR sequences, UTR start and UTR end coordinates of a specific gene in rat.
  - Retrieve entry for an accession BC037153.
  - Get protein sequence for NP_000305.3.

File formats

- Examples ?

File formats

- Fasta
- BED
- Genbank format
- Swissprot/TrEMBL format

Format conversion is the key

Yes, we keep doing that.

How can we manipulate data/files/formats ?

Break